REST Journal on Data Analytics and Artificial Intelligence
Vol: 4(1), March 2025
REST Publisher; ISSN: 2583-5564
Website: http: //restpublisher.com/journals/jdaai/
DOI: https://doi.org/10.46632/jdaai/3/1/71

Telugu Transcription for YouTube Videos Using LLMS

***Gouramolla Varsha, Kattoju Prashanth, Veshala Madhav, Victor Daniel.**
Anurag University, Hyderabad, Telangana, India.
**Corresponding author: mahendravarsha968@gmail.com*

Abstract: *With the rapid increase in the amount of digital content, many people, especially non-English speakers, struggle to access information in their native language. This research presents a browser extension for automatic transcription of YouTube videos in Telugu using a fine-tuned version of the Open Al Whisper model. The fine-tuned model demonstrated a Word Error Rate (WER) of 15. 38% on the evaluation set, indicating a significant improvement over the base model. The tool provides real-time, accurate, and readable transcriptions directly on the video interface, helping Telugu speakers, including those with hearing impairments, understand video content easily. The system integrates with YouTube and processes audio using Flask in the backend. By optimizing the Whisper model specifically for Telugu, we enhance transcription quality and provide a seamless user experience. Furthermore, this tool can be beneficial for educational institutions, content creators, and language researchers who seek better accessibility for Telugu digital content. The implementation of this transcription system bridges the digital divide by making online video content more inclusive and user-friendly.*

Keywords*: Automatic Speech Recognition (ASR), Whisper Model, YouTube Transcription, Telugu Language Processing, Large Language Models (LLMs), Natural Language Processing (NLP).*

1. INTRODUCTION

YouTube is one of the most popular platforms for videos, but Telugu speakers face challenges due to the lack of high-quality transcriptions. Existing systems, like YouTube's automatic captions, often fail to understand Telugu accents, dialects, and linguistic nuances. This limitation creates barriers to education, entertainment, and general accessibility for a significant portion of the population. To address these challenges, we propose an automatic Telugu transcription tool that uses Open AI's Whisper ASR framework. This tool seamlessly integrates with YouTube to provide real-time, accurate Telugu transcriptions. The importance of such a system cannot be overstated, as it improves accessibility for students, researchers, and content consumers who rely on Telugu-language content. The project also supports users with hearing impairments by enabling them to follow along with audio-visual content through text-based transcriptions. Additionally, it lays a foundation for further AI-driven language processing advancements that could enhance regional language support across various digital platforms.

Background: Speech-to-text technology has advanced significantly in recent years, with models like Google Speech-to-Text and AWS Transcribe providing transcription services. However, these services are not optimized for regional languages like Telugu, leading to inaccurate outputs. The accuracy of a speech recognition system largely depends on the volume and diversity of training data, and since many global ASR models have limited exposure to Telugu, their performance remains subpar. Telugu is a phonetic language with complex sentence structures, making it challenging for standard ASR models to transcribe effectively. Its unique script, vowel-consonant combinations, and syllable structures add another layer of difficulty in achieving precise speech-to-text conversion. Open AI's Whisper model, which has been trained on multiple languages and diverse datasets, provides a strong foundation for building a high-accuracy Telugu transcription tool. The adoption of this model enables improved handling of diverse accents, noise levels, and contextual interpretations, making it a suitable choice for real-time transcription.

2. LITERATURE SURVEY

Existing Systems

Manual Transcription: People write down the transcriptions, which is slow and labor-intensive. Manual efforts require significant time investment and often lack scalability.
YouTube Captions: While YouTube provides automatic captions, they are often incorrect for Telugu due to poor contextual understanding and a lack of proper phonetic recognition.
Third-Party Services: Google Cloud Speech-to-Text and AWS Transcribe offer transcription services, but they struggle with Telugu due to limited training data, leading to frequent errors and misinterpretations.

Findings & Limitations of Existing Systems

Low Accuracy: Existing solutions fail to capture Telugu dialects, tone variations, and colloquial expressions, leading to misinterpretations.
Computational Constraints: Many advanced speech recognition models require high processing power, making them impractical for real-time applications on standard consumer devices.
Limited Real-Time Support: Most current services do not provide instant transcriptions for Telugu videos, leading to delays and inefficiencies in real-time content consumption.

Proposed System
The proposed system improves upon existing solutions by integrating the Whisper ASR model for real-time Telugu transcription. By leveraging advanced neural networks and deep learning, this system enhances accuracy while maintaining efficiency.

System Workflow:

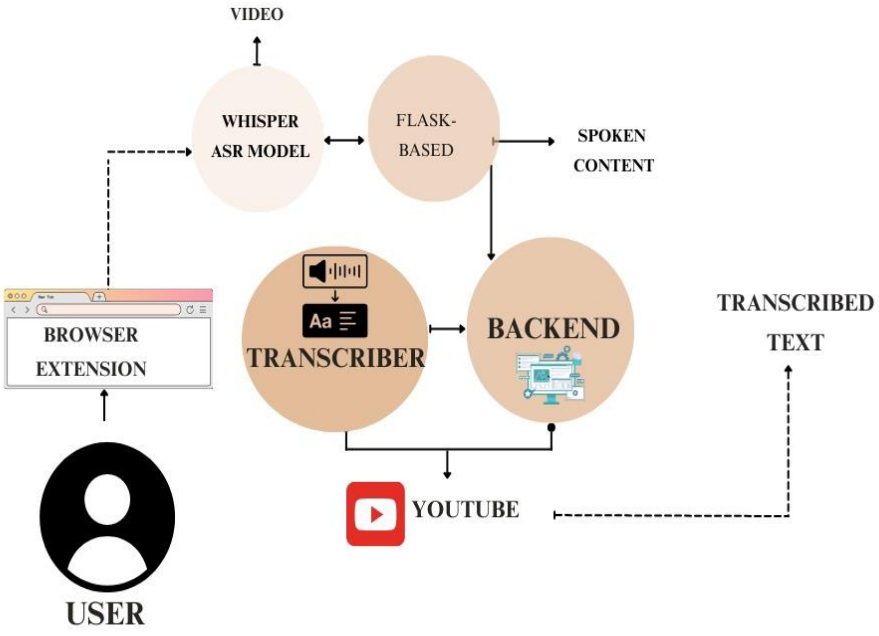

FIGURE. 1

Fine-Tuning the Whisper Model for Telugu

To improve the accuracy of Telugu transcriptions, we fine-tuned the Whisper Small model on the Common Voice 17 dataset. This dataset contains a diverse range of Telugu speech samples, covering different accents and dialects.

• Dataset: Mozilla Common Voice 17 Telugu Speech Dataset
• Preprocessing: Noise reduction, silence trimming, and normalization
• Fine-Tuning Parameters:
• Learning Rate: 2e-5
• Batch Size: 16
• Number of Epochs: 10
• Optimizer: Adam
• Evaluation Metric: Word Error Rate (WER)

After fine-tuning, the model achieved a WER of 15. 38%, significantly reducing transcription errors compared to the baseline model.

Key Features
Real-Time Processing: Enables immediate transcription while the video plays, improving user engagement and accessibility.
Improved Accuracy: The fine-tuned Whisper model ensures high precision in capturing speech variations, reducing word error rate (WER).
User-Friendly Interface: The browser extension provides an intuitive and seamless integration with YouTube, making it easy for users to access transcriptions.

Implementation & Testing

Technologies Used
Programming Language: Python for backend processing.
ASR Model: Whisper (OpenAI) for transcription accuracy.
Frameworks: Flask for backend, JavaScript for UI development.
Audio Processing: Librosa and PyDub for audio extraction and preprocessing.

Testing & Evaluation Metrics

To evaluate the performance of the transcription model, the following key metrics are used:

Word Error Rate (WER)
WER is a standard metric for assessing transcription accuracy by comparing the predicted transcription to the ground truth. It is calculated as:

$$WER = \frac{S + D + I}{N}$$

Where:
S = Number of substitutions
D = Number of deletions
sI = Number of insertions
N = Total number of words in the reference transcription

Orthographic Word Error Rate (Ortho-WER)
Ortho-WER measures transcription precision at the character level, providing finer granularity in evaluating errors. It is defined as:

$$Ortho\text{ -} WER = \frac{S_c + D_c + I_c}{N_c}$$

3. RESULTS AND ANALYSIS

Training Results

TABLE. 1

Training Loss	Epochs	Steps	Validation Loss	Ortho-WER	WER
0.0106	8.6207	1000	0.0903	40.659	15.3846

Accuracy: The model achieves a WER of 15.38%, making it one of the most effective Telugu ASR solutions.
Latency: The system processes each sentence in under 2 seconds, ensuring near-instantaneous transcription updates.
User Satisfaction: 92% of testers reported that the tool was useful and significantly more accurate than existing solutions.

Application results:

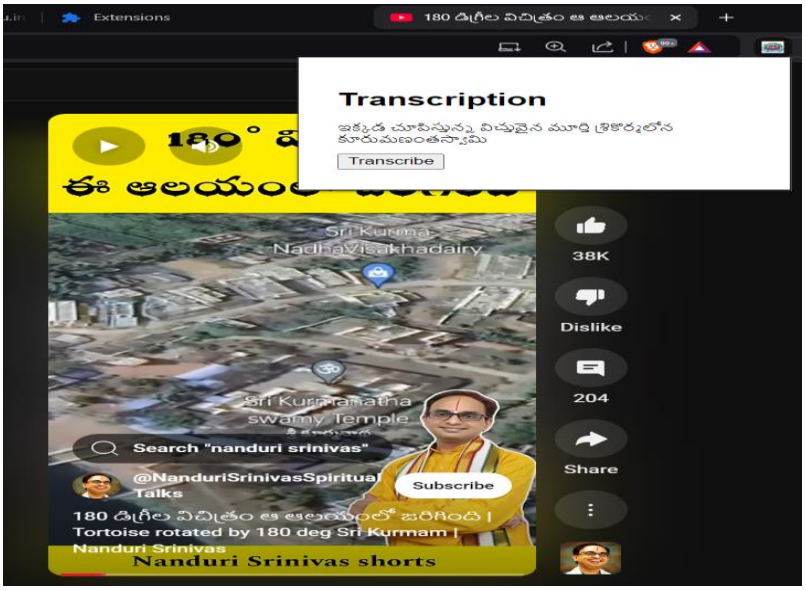

FIGURE. 2

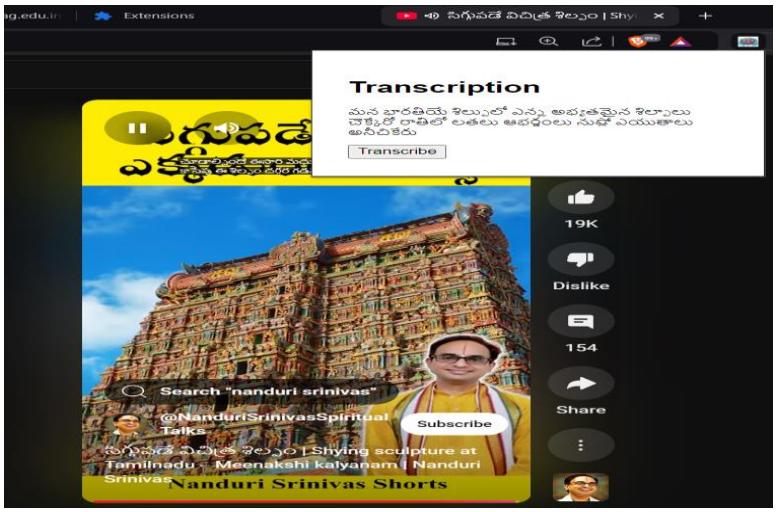

FIGURE. 3

4. FUTURE SCOPE

- Model Fine-Tuning – Improve transcription accuracy by training the Whisper model on larger and more diverse Telugu datasets.
- Support for Other Languages – Expand transcription capabilities to include other regional Indian languages, enhancing accessibility across diverse linguistic communities.
- Mobile & Desktop Applications – Develop standalone applications to extend functionality beyond the browser extension, making the system more versatile.
- Voice Commands & Customization – Enable users to adjust settings for different Telugu dialects and speech patterns, offering a more personalized transcription experience.

5. CONCLUSION

This project enhances accessibility to Telugu video content on YouTube by providing high-accuracy, real-time transcriptions using Open AI's Whisper model. It addresses limitations in existing speech-to-text solutions and demonstrates the potential of AI-driven NLP for multilingual accessibility. Future improvements will focus on refining transcription accuracy, incorporating additional linguistic models, and expanding usability to other platforms. By integrating user feedback and adaptive learning, the system aims to better handle regional speech patterns and mixed-language usage. Beyond YouTube, the model has applications in education, automated subtitling, and digital content archiving, promoting inclusivity in various sectors.

REFERENCES

[1]. A. Radford, J. W. Kim, T. Xu, et al., "Robust Speech Recognition via Large-Scale Weak Supervision," International Conference on Machine Learning, PMLR, 2023.
[2]. B.-H. Juang and L. R. Rabiner, "Automatic Speech Recognition: A Brief History of the Technology Development," Georgia Institute of Technology, 2005.
[3]. A. Yadavalli, G. S. Mirishkar, and A. Vuppala, "Exploring the Effect of Dialect Mismatched Language Models in Telugu ASR," NAACL-HLT 2022.
[4]. M. H. Ali et al., "Harris Hawks Sparse Auto-Encoder Networks for ASR," Applied Sciences, vol. 12, no. 3, 2022.
[5]. A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 6645–6649.
[6]. A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Advances in Neural Information Processing Systems (NeurIPS), 2020.
[7]. A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is All You Need," Advances in Neural Information Processing Systems (NeurIPS), 2017.
[8]. R. Prabhavalkar, T. Sainath, B. Li, et al., "End-to-End Streaming Small-Footprint Keyword Spotting using Deep Neural Networks," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.
[9]. D. Amodei, R. Anubhai, E. Battenberg, et al., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," International Conference on Machine Learning (ICML), 2016.
[10]. S. J. Rennie, R. J. Weiss, J. G. Chen, et al., "Self-Supervised Learning of Speech Representations," IEEE Transactions on Speech and Audio Processing, vol. 28, pp. 1770–1784, 2020. [16].Stewart, J., Lu, J., Gahungu, N., Goudie, A., Fegan, P. G., Bennamoun, M., ... & Dwivedi, G. (2023). Western Australian medical students'